Multimedia Event Extraction (M^2E^2) Annotation Guideline V0.1

Manling Li[1]*, **Alireza Zareian**[2]*, **Qi Zeng**[1], **Spencer Whitehead**[1], **Di Lu**[3], **Heng Ji**[1,4], **Shih-Fu Chang**[2,4]
[1]University of Illinois at Urbana-Champaign
[2]Columbia University [3]Dataminr [4]Amazon Alexa AI
hengji@illinois.edu,sc250@columbia.edu

1 Introduction

In this paper we propose a multimedia event extraction (M^2E^2) task, where the input is a text news article with images, and the output is the events extracted from text and image modalities.

We construct the M^2E^2dataset [1] by collecting all of the 108,693 multimedia news articles from the Voice of America (VOA) website [2] 2006-2017 and selecting 245 multimedia documents as the annotation set based on three criteria: (1) Informativeness: prefer articles with more event mentions; (2) Illustration: prefer articles with more images (at least four); (3) Diversity: balance the event type distribution. For the first and third criteria, we use the baseline text-only event extraction model (Li et al., 2019) to estimate the number of event mentions per event type in each articles.

Modality	Type	Task
Text	Event Type	Classification (Event Type)
		Localization (Trigger)
	Argument	Classification (Argument)
		Localization (Entity)
Image	Event Type	Classification (Event Type)
	Argument	Classification (Argument)
		Localization (Union)
		Localization (Instance)
Cross	Coreference	Classification (Relation)

Table 1: M^2E^2 annotation tasks.

We annotate event type and argument roles for textual and visual events. The annotation process involves tasks in Table 1. After completing text-independent and image-independent annotations, expert annotators are asked to perform adjudication. We do **not** tag all events, but only a particular subset from ACE ontology, as in Table 2.

Textual event annotation includes event type annotation and argument annotation. we assign an event type to each event trigger (the words or phrases that most clearly express event occurrences), and an argument role to each participant (entity, time or value expression). Here we focus on intra-sentence event extraction and do not consider cross-sentence or cross-document situations.

Visual event annotation includes event type annotation and argument annotation. We assign an event type to each image if the image contains predefined types of events, and assign argument roles to corresponding bounding boxes. The event type annotation does not locate a specific region in the image, but use the whole image as justification.

After annotating events and arguments separately in each modality, we ask annotators to find image-sentence pairs that correspond to the same event instance, i.e., the same event type happening in the same location and time.

This guideline focuses on how to annotate events and argument roles in images. For more details in text event annotation, please refer to the *ACE English Annotation Guidelines for Events (Version 5.4.3 2005.07.01)*[3].

2 Image Event Type Annotation

Caption as reference. The image caption is used as reference when deciding the image event type. It may be ambiguous to determine event type only by the image itself, as examples in Figure 1, Figure 2 and Figure 3 show. If the image does not show some actions in particular tagged events from the caption, as in 5, it will not be tagged either. However, when the image implies an event that is not in the caption, we will annotate the image. Take Figure 4 as an example.

*These authors contributed equally to this work.

[1] http://blender.cs.illinois.edu/software/m2e2

[2] https://www.voanews.com/

[3] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

Event Type	Definition
Life.Die	it occurs when the life of a person entity ends
Movement.Transport	it occurs when an *artifact* or a *person* is moved from one place to another
Transaction.TransferMoney	it refers to the giving, receiving, borrowing, or lending money
Conflict.Attack	a violent physical act causing harm or damage
Conflict.Demonstrate	it occurs when a large number of people come together in a public area to protest or demand some sort of official action
Contact.Meet	it occurs when two or more people interact with one another face-to-face at a single location
Contact.Phone-Write	it occurs when two or more people directly engage in discussion but not face-to-face
Justice.ArrestJail	it occurs when the movement of a person is constrained by a state actor

Table 2: Event type definition in $\mathbf{M^2E^2}$dataset.

Event Type	Argument Role	Argument Role Definition
Life.Die	Agent, Victim, Instrument, Place	[Victim] died at [Plcae], or was killed by [Agent] using [Instrument]
Movement.Transport	Destination, Origin, Instrument, Agent, Artifact/Person	[Agent] transported [Artifact or Person] in [Instrument] from [Origin] to [Destination]
Transaction.TransferMoney	Giver, Recipient, Money,	[Giver] gave [Money] to [Recipient]
Conflict.Attack	Attacker, Instrument, Place, Target	[Attacker] attacked or assaulted [Target] using [Instrument] at [Place]
Conflict.Demonstrate	Demonstrator, **Instrument**, **Police**, Place	[Demonstrator] was in a demonstration at [Place] holding [Instrument] and supervised by [Police]
Contact.Meet	Participant, Place	[Participant] met face-to-face at [Place]
Contact.Phone-Write	Participant, Instrument, Place	[Participant] communicated with [Participant] at [Place] (not explicitly face-to-face)
Justice.ArrestJail	Agent, Person, **Instrument**, Place	[Agent] arrested or jailed [Person] using [Instrument] at [Place]

Table 3: Event arguments in $\mathbf{M^2E^2}$dataset. Extended arguments are in bold.

Figure 1: *Life.Die*: incorrect event type annotation. It should be *Movement.transport*. Image caption: *After crossing from Serbia into Asotthalom, Hungary, a man cradles a child and waits with other migrants for* ***transfer*** *to a refugee camp, Aug. 31, 2015.*

Multiple events in one image. For one image indicating multiple events, such as Figure 4, it will be tagged with multiple event types. For one caption indicating multiple events, such as Figure 7, we will only tag the image with multiple event types when the images do indicate them.

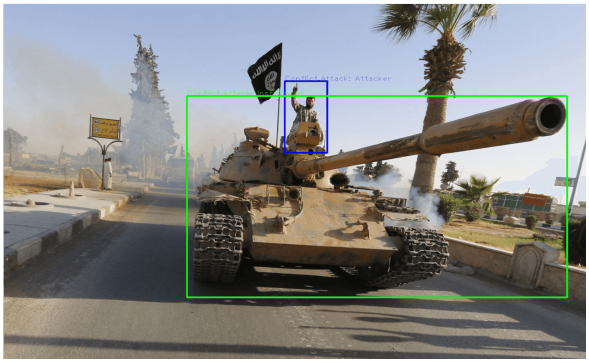

Figure 2: Conflict.Attack: incorrect event type annotation. Image caption: *Militant Islamist* ***fighters*** *on a tank take part in a military parade along the streets of northern Raqqa province, June 30, 2014.*

3 Image Event Argument Annotation

Union and Instance Bounding Boxes. We adopt bounding box, the smallest region covering the target, to label event arguments. Bounding boxes of two granularity are used, union-level and instance-level. As shown in Figure 8 , the union bounding box for each role is a smallest bounding box covering all the arguments. In comparison, the instance bounding box specifies each visual object instance, i.e., one argument. In Figure 9, each role has multiple instance bounding boxes. The cri-

Figure 3: Conflict.Demonstrate: incorrect event type annotation. Image caption: *People **celebrate** Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington, D.C., June 26, 2015.*

Figure 4: *Contact.Meet* and *Contact.Phone-Write*: this image indicates two events. Image caption: Belgian Prime Minister Charles Michel, center, addresses Belgium's parliament announcing security measures after the recent deadly Paris attacks, in Brussels, Nov. 19, 2015.

Figure 5: *Untaggable:* although the image caption directs to event type of *Life.Die*, this image is untaggable because expected actions or arguments do not show up in the image . Image caption: *A Pakistani woman mourns the death of her family member outside a mortuary in Karachi, Pakistan, September 12, 2012.*

Figure 6: *Untaggable:* the image caption does not imply any event type. Image caption: *A man suffering from the Ebola virus lies on the floor outside a house in Port Loko Community, situated on the outskirts of Freetown, in Sierra Leone, Oct. 21, 2014.*

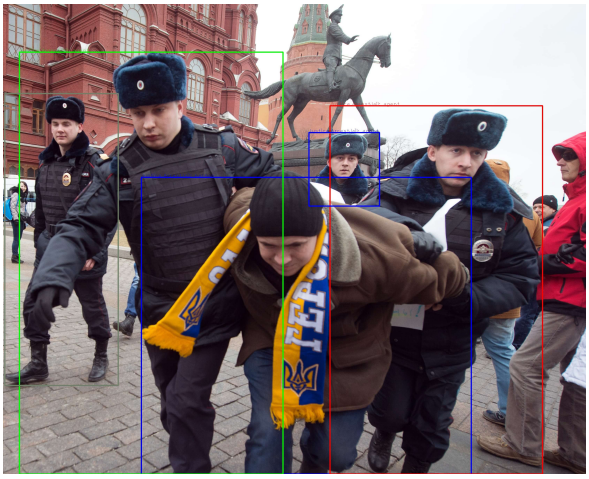

Figure 7: *Justice.ArrestJail*: although the word "protester" implies a *Conflict.Demonstrate* event, we only tag this image as *Justice.ArrestJail*. Image caption: *Police officers **detain** a **protester** in central Manezhnaya Square in Moscow, on March 2, 2014, during an unsanctioned rally against the Russia's military actions in Crimea.*

teria of instance bounding box follows the visual object annotation guideline *VOC 2011 Annotation Guidelines*[4].

Extended Roles. Based on the observation that some visual arguments are usually not likely to show up in the text, we extend the argument list for some event types from ACE annotation guideline. For example, "Instrument" in *Conflict.Demonstrate* events, usually the poster or board, has a much lower frequency in text than in image but provide much information.

Skipped Roles. Some argument roles, such as

[4] http://host.robots.ox.ac.uk/pascal/VOC/voc2011/guidelines.html

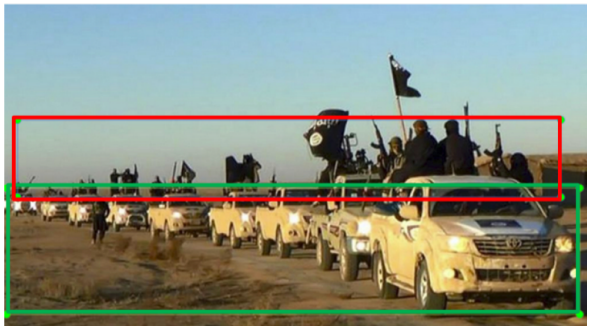

Figure 8: Example of a union bounding box.

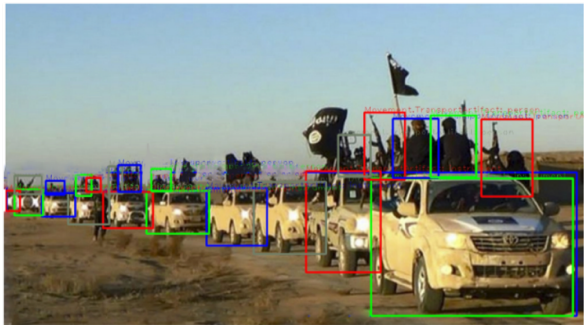

Figure 9: Example of instance bounding boxes.

"Place", are too abstract to be an instance in images. Therefore we do not label those vague argument roles in image event argument annotation.

Figure 10: *Movement.Transport*: Instrument (green), Person (blue, **incorrect** annotation as the other two refugees should be also included), agent (red, incorrect annotation) Image caption: *Syrian refugees lie exhausted moments after arriving by a raft at a beach on the Greek island of Lesbos, Oct. 25, 2015.*

References

Manling Li, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. 2019. Multilingual entity, relation, event and human value extraction. In *NAACL 2019*, pages 110–115.

Figure 11: *Transaction.TransferMoney*: Money (blue), Recipient (green and red). Image caption: *Paul White (L), 45, from Ham Lake, Minnesota, stands with his partner Kim VanReese (C) and co-worker Nancy Bowen (R) as he holds a check for his $149.4 million portion of a $448.4 million Powerball jackpot prize at a news conference at Minnesota State Lot*

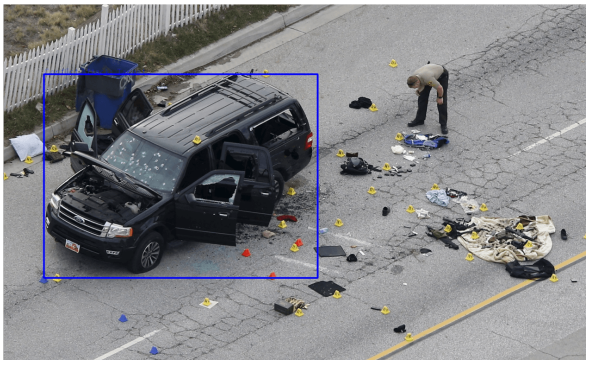

Figure 12: *Conflict.Attack*: Target (blue). Image caption: *A law enforcement officer looks over the evidence near the remains of a **SUV** involved in the Wednesdays **attack** is shown in San Bernardino, California, Dec. 3, 2015.*

Figure 13: *Conflict.Attack*: Target (blue, **incorrect** annotation as only the SUV is attacked based on the text). Image caption: *A police officer picks up a weapon from the scene of the investigation around the area of the SUV vehicle where two suspects were shot by police following a mass **shooting** in San Bernardino, California, Dec. 3, 2015.*

Figure 14: Conflict.Demonstrate: Demonstrator (blue, **incorrect** annotation as the demonstrators are not labeled in a minimal bounding box) Image caption: *Activists deliver more than 400,000* ***petition signatures*** *to Capitol Hill in support of the Iran nuclear deal in Washington, July 29, 2015.*

Figure 15: *Conflict.Demonstrate*: Demonstrator (blue) and Instrument (green). Image caption: *Kurdish people gather in front of the building where three Kurdish women were killed in Paris, France, January 10, 2013.*

Figure 16: *Contact.Meet*: Participant (blue and green, **incorrect** annotation because the woman is also Participant). Image caption: *U.S. Secretary of State John Kerry* ***meets*** *with Iraq's Foreign Minister Ibrahim al-Jaafari (R) in Baghdad September 10, 2014. Kerry identified Iraq as a key partner in the fight against IS. REUTERS/Thaier Al-Sudani*

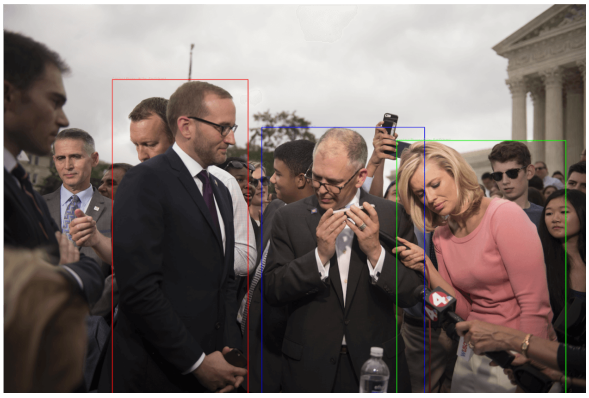

Figure 17: *Contact.Phone-Write*: Participant. **Incorrect** annotation because the cell phone should be annotated as Instrument). Image caption: *Jim Obergefell, the named plaintiff in the case before the Supreme Court, center,* ***talks on a cellphone*** *to President Barack Obama on the steps of the Supreme Court following the court's decision, in Washington, D.C., June 26, 2015.*

Figure 18: *Contact.Phone-Write*: Participant (green) and Instrument (blue). Image caption: *A Bangladeshi student* ***talks*** *on a mobile phone at Dhaka University campus, July 2001.*